DOCUMENT RESUME

ED 167 620 TM 008 434

AUTHOR White, Owen R.
TITLE Curriculum-References, Time-Series Program Evaluation.
PUB DATE Mar 78
NOTE 25p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Case Studies; Control Groups; Elementary Secondary Education; Evaluation Methods; *Exceptional Child Services; Handicapped Children; *Individualized Programs; Norm Referenced Tests; *Program Evaluation; *Research Design; *Special Education; Testing; Test Validity
IDENTIFIERS *Time Series Analysis

ABSTRACT

Traditionally, the use of control groups, regression discontinuity/projection, or norm referenced models has been advocated for educational program evaluation. The small numbers of and the heterogeneous characteristics of severely and profoundly handicapped make these evaluation methods inappropriate for this population subgroup. Furthermore, no overall group summary of pupil progress will meet the legislative mandates in recent laws (PL 93-380 and PL 94-142) for the evaluation of individual programs. The most commonly adopted alternative, intensive or single subject evaluation, also presents problems because it is often difficult to summarize information across single subject evaluations to form a meaningful picture of overall program impact. Curriculum-referenced evaluation is a relatively new approach, based on the notion that a program's impact should be measured by compiling a list of objectives to be checked off as they are mastered. Since the curriculum is program specific, such assessments have greater validity than those based on national norms. Meaningful summarization across children and program elements is possible. Believable evidence of program impact may then be obtained by conducting formal or informal time--series evaluations during or following program involvement. Several instruments currently exist for special education programs, for example, the University of Washington's Uniform Performance Assessment System, Oregon's Student Progress System, and the Teaching Research Curriculum. Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CURRICULUM-REFERENCES, TIME-SERIES PROGRAM EVALUATION

Owen R. White

Experimental Education Unit, College of Education
Child Development and Mental Retardation Center
University of Washington

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Owen R. White

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED167620

Introduction

The number of evaluation and research designs which might be brought to bear on the issue of educational program effectiveness and generalizability can virtually stagger the mind. Practically speaking, however, it is often impossible or highly impractical to utilize those designs which have been traditionally held in the highest esteem for their power and procesion. Many would choose to simply avoid those problems by recommending that programs which cannot mount highly sophisticated and expensive evaluation operations simply accept the fact that they cannot meaningfully evaluate their effectiveness (e.g., Cooley & Lohnes, 1976). In the face of increasing legislative mandate (e.g., United States Public Laws 93-380 and 94-142), that is no longer a tenable position. Virtually all educational programs must make some attempt to demonstrate their effectiveness and, at least with handicapped children, that evaluation must reach down to the level of the individual child. The time for chastising programs for their inability to utilize traditional evaluation strategies is past. It is now time to face the realities of the situations in which evaluations must take place and to identify reasonable evaluation strategies which have at least some potential for yielding meaningful results within the practical constraints of those situations. It shall be the purpose of this presentation to at least initiate that process.

Practically Speaking

From the standpoint of evaluative need, educational programs might be divided into two categories: (1) special, heavily funded demonstration or research programs (where, presumably, rather special resources might be brought to bear on the issue of evaluation); and (2) ongoing service programs. For purposes of these discussions, I shall concentrate on the latter. Moreover, since they present some very special evaluation problems, I shall concentrate on programs serving the severely and profoundly handicapped. It should be noted, however, that in many cases, the constraints which are imposed upon such programs are also encountered in other programs.

TM008 434

When designing an evaluation strategy for ongoing educationa service programs, at least the following factors must be considered:

(1) Personnel and Other Resources. Most evaluations must be conducted with regular personnel. Such personnel are not likely to be highly trained in evaluation or research strategies, nor are they likely to have a great deal of extra time to devote to program evaluation--especially if the results of those evaluations are not organized in a way which is of immediate practical usefulness to them in their roles as educators.

(2) Ethical Group Selection. Few programs have any real choice over the selection of the pupils they will serve. At best, certain criteria for acceptance will be established so that all children in a given program share some characteristics in common. In most cases, however, it is simply not possible to fulfill the requirements of a powerful program evaluation strategy by preventing children from entering certain programs which might best meet their needs.

(3) The Need to Improve Programs. Most educational programs will not be interested in validating an entire service "package" for dissemination to other interested parties. Rather, emphasis must be placed on evaluation strategies which contribute to the continuous refinement and improvement of the program--preferably at a level where individual teachers and children currently in the system will demonstrably realize the benefits of those improvements. Strict timelines defining periods of preplanned program manipulations will not, therefore, be realistic. The evaluation program must be flexible enough to allow the utilization of information in the refinement of individual and group programs as soon as that information is available.

(4) Individual Evaluations. As mentioned earlier, laws now exist which require certain programs to measure the impact of their efforts at the level of the individual child. At least with programs serving handicapped children, that means more than simply recording the results of individual efforts on group-referenced achievement tests. It means the systematic appraisal of progress toward objectives established individually for each child in what the law (P.L. 94-142) terms the Individualized Educational Plan. At very least, overall program evaluation plans must consider the fact that some teacher time must be spent in the evaluation of individual progress (i.e., time which might otherwise be spent in the evaluation of overall group progress) and, wherever possible, some attempt should be made to save teacher time by finding ways of meaningfully incorporating those individualized data into the overall program evaluation.

(5) The Severely and Profoundly Handicapped. Several special problems are encountered when attempting to evaluate programs serving the severely and profoundly handicapped. First, they

represent a very low-density subsection of the general population. A school district with 20,000 children, for example, might expect to serve only 150 to 200 such children between three and twenty-one years of age (i.e., the age range mandated by law for serving the severely and profoundly handicapped). With so few children, there is very likely only one program in any given district or (for rural communities) geographic region, thereby limiting the availability of "contrast" or "control" groups which might be used in program evaluations. The sparsity of programs also exacerbates the problem of subject selection--virtually all severely and profoundly handicapped children in a given area must be placed in one program, regardless of their specific disabilities or personal characteristics. There is simply no other place to go. That raises the second major evaluation problem: the subpopulation of severely and profoundly handicapped children is extremely heterogeneous. There are eight major handicapping conditions recognized by the Bureau of Education for the Handicapped. Most severely and profoundly handicapped children are handicapped in more than one way, however, raising the number of combined categories to 46--without considering level of deficit or age. In the hypothetical district of 20,000, then, it is quite unlikely that more than two or three severely or profoundly handicapped children could be found which could be reasonably matched on the basis of even the most obvious educationally relevant variables. As will be demonstrated below, the low density and high heterogeniety characteristics of the severely and profoundly handicapped subpopulation of children virtually nullifies the possibility of utilizing the vast majority of traditional evaluation strategies with programs serving such children.

Problems with Group Designs

The use of equivalent control groups has traditionally been favored for the evaluation of educational programs. In those programs where it is not possible or practical to estabilsh directly equivalent control groups, systematically different controls might be useful with regression-projection or regression-discontinuity analyses. If neither of those two approaches is possible, it has been suggested that meaningful evaluations might still be conducted through the careful use of norm-referenced assessment procedures (Tallmadge & Horst, 1976). For many programs, especially those serving the severely and profoundly handicapped, there are problems associated with all of those approaches.

Equivalent Control Groups

When dealing with programs serving normal children or mildly handicapped children, many districts would find themselves with a sufficient subject population to warrant the consideration of one or more types of

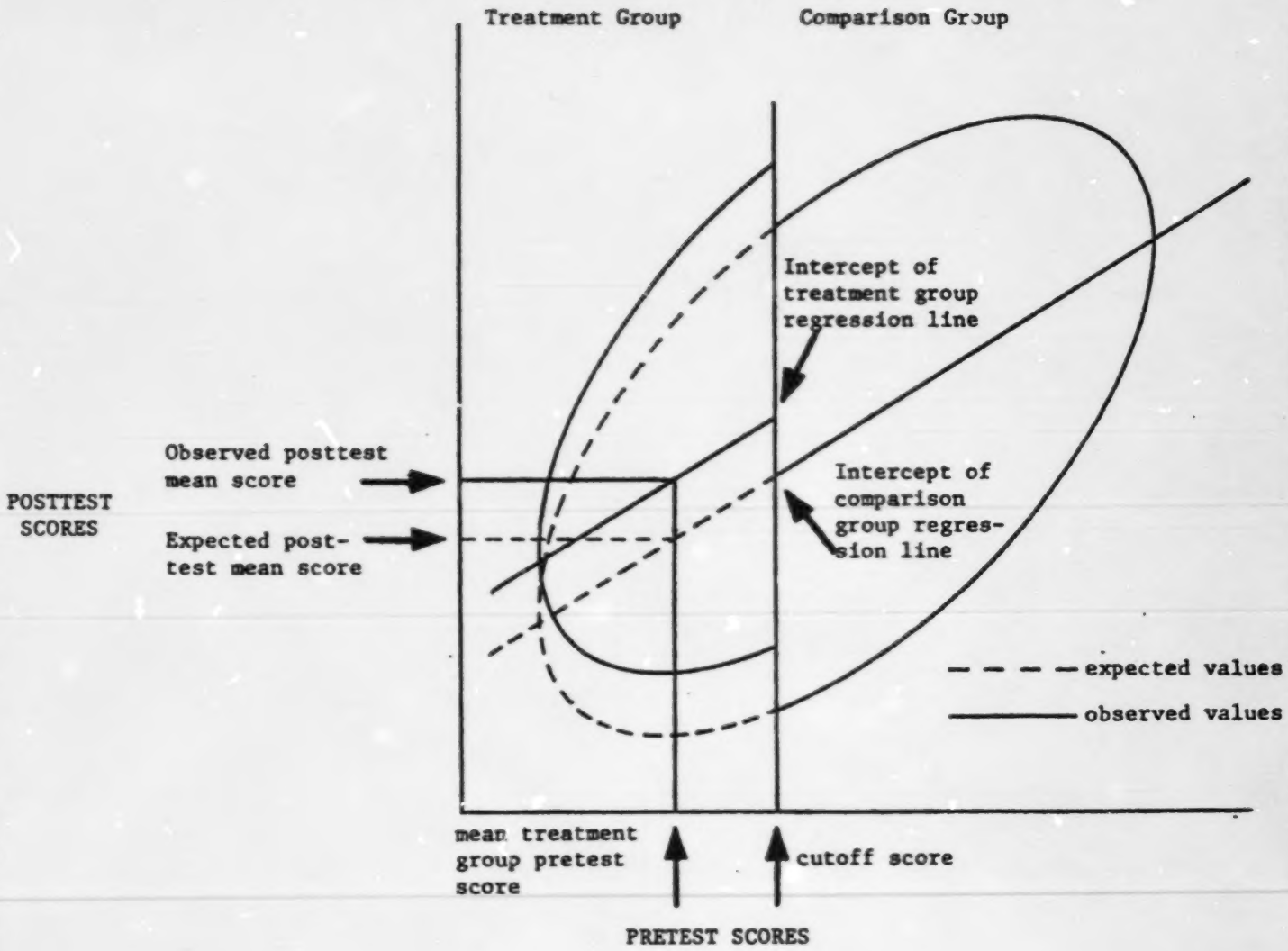

Figure 1: The Regression-Discontinuity Model

Adapted from Horst, Tallmadge & Wood (pg. 61, 1975)

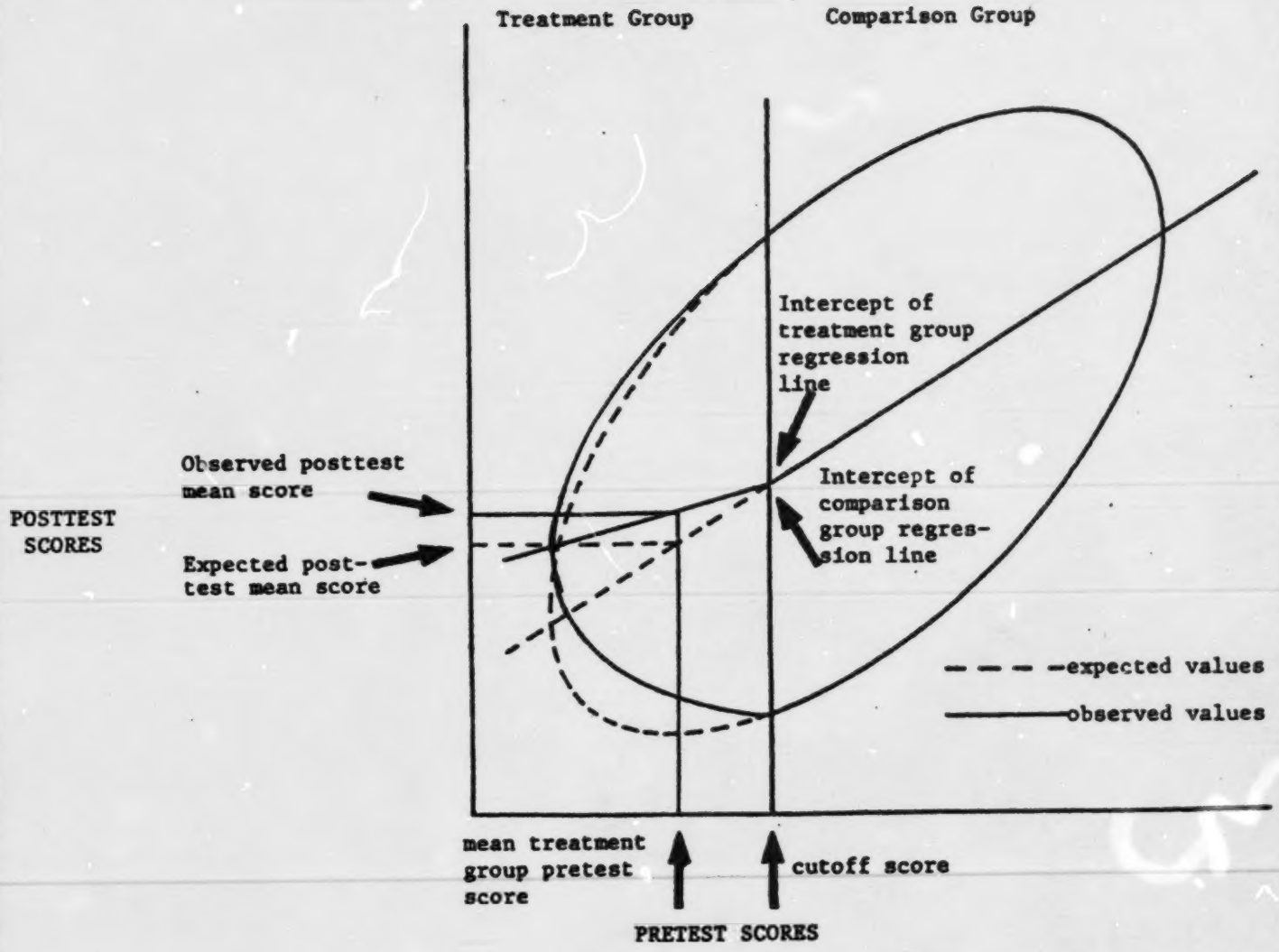

Figure 2: The Regression-Projection Model

Adapted from Horst, Tallmadge & Wood (pg. 62, 1975)

equivalent control group evaluation designs--assuming, of course, that there were no ethical problems in assigning children to different treatment of control groups. However, much of the emphasis for evaluation in education today is being directed at "special" programs serving subsections of the pupil population. If the target group is small, then the program in question may be the "only game in town," obviating direct comparisons between service options already in existance at different sites within the same general service area. Locating programs in other service areas to serve as controls raises issues of "natural selection" and, from the practical standpoint, political problems arise in establishing a cooperative assessment program designed to demonstrate that one district's program is superior to another district's program. The obvious solution is to assign children within the same program to two or more different treatment groups. That will only be a viable option, however, if the programs serve a large number of pupils and/or the children in the program represent a relatively homogeneous group with respect to most educationally relevant variables. As mentioned earlier, neither set of conditions is likely to obtain with programs serving at least some types of pupils (e.g., the severely and profoundly handicapped).

Regression-Projection and Regression-Discontinuity Models

When pupils are selected for inclusion in a program on the basis of some absolute cutoff score on a pretest assessment, it may not be necessary to find directly equivalent control groups for program evaluation. The relationship between pretest and posttest scores for children not included in the program (i.e., those performing above the cutoff level on the test) might be used to formulate regressions which estimate the expected relationship between pretest and posttest socres for children included in the program. Actual posttest scores of the treated group are then compared against that estimate to evaluate program impact. If the impact is likely to be the same for all children, regardless of entry score, then the regression-discontinuity model is appropriate (Campbell & Stanley, 1963). If the treatment is likely to differentially affect pupil gains in accord with entry performance, then the regression-projection model is more appropriate (Tallmadge & Horse, 1974) (see Figures 1 and 2). In either case, it is assumed that the performance gains of higher-level children can be used to estimate the expected performance gains of lower-level children. That may be an unwarranted assumption.

- - - - - - - - - - - - - - - - - -
Insert Figures 1 & 2 about here
- - - - - - - - - - - - - - - - - -

In order to assume that the relationship between initial performance level and performance gain is predictable and "smooth" across all levels of performance, one must assume that the dependent measures are a function of some set of attributes or conditions which are also distributed smoothly throughout the range of concern. While some characteristics of a population might be distributed in that manner (e.g., "intell igence"), others are definitely not. For example, in measuring the

impact of a program on the development of fine motor movement in crippled children, it would be foolish to assume that the relationship of initial performance and performance gain in non-crippled children was at all similar. The same might also be said for the growth characteristics of blind children, deaf children, emotionally disturbed children, or even bilingual children. Obviously, there are many types of educational programs where the regression-projection and regression-discontinuity models cannot be used with confidence.

Norm-Referenced Models

In cases where no empirical data can be gathered from a reasonable control or contrast group concurrently with the data gathered from the treatment group of concern, it has been suggested that the use of norm-referenced tests can be of value (Tallmadge, 1977). In essence, one uses data gathered by other persons on other children as the "control." However, as Horst and Tallmadge (1975) point out:

> The validity of the (norm-refernced) models rests on the assumption that the achievement status of a particular subgroup remains constant relative to the norm group over the pre- to post-test interval if no special treatment is provided. Empirical support for this assumption is minimal. It is conceivable that some subgroups would move up and others move down in the normal course of events. When the norm group is like the treatment group, the plausibility of the underlying assumption is greatly enhanced; thus, for example, norms for gifted children would be best for assessing a project serving such pupils. (p. 72)

Unfortunately, special norms for selected subsections of the general pupil population are virtually nonexistent, especially for the severely and profoundly handicapped. The validity of existing norms for such children (and other types of pupil subpopulations) are in even greater doubt than with most subgroups of the population since few, if any, of those children were included in the norming samples for most instruments. As a result, the authors and developers of the most widely used perfomrnace assessment instruments usually warn the user not to employ the test with children suspected of severe handicapping conditions, and most raw-score conversion tables do not include values suitable for such children. Attempts are being made to develop performance norms for handicapped children, but due to the low-density nature of the severely/profoundly handicapped subpopulation, even the most extensive effort to date (Sailor & Mix, 1975) was only able to include some 300 pupils in its norming sample--pupils with all types of handicapping conditions and ranging from three to 21 years of age!. Obviously, valid norm-refernced evaluation models for at least some types of educational programs remain only a dim hope for the future.

And There's Still the Individual...

Beyond the problems associated with control groups, regression-discontinuity/projection and norm-referenced models in particular, there

is one major problem with most traditional group approaches: education is becoming increasingly concerned with individual pupil progress. No overall group summary of pupil progress will meet the legislative mandates in recent laws (P.L. 93-380 and P.L. 94-142) for the evaluation of individual programs. The value of group summaries cannot be underestimated for purposes of general administrative decision-making, but some method must be devised which allows such summaries to be compiled without overlooking the individual nature of each pupil's needs and program.

Problems with "Intensive" or "Single-Subject" Evaluation Designs

In applied behavior analysis, the use of individual subjects as their own controls in time-series investigations has been a favorite approach for some time (e.g., Sidman, 1960; Hernsen & Barlow, 1976). Basically, repeated assessments are conducted of an individual pupil's performances under both "treatment" and "non-treatment" conditions. Levels of performance and rates of progress under the two conditions are then compared to determine the impact of the treatment. Such an approach certainly offers many advantages, especially from the standpoint of meeting the conern for individual progress. Still, the single-subject approach is no panacea.

Basic Approaches

There are three basic approaches to single-subject evaluation: the "withdrawal" or "ABA" design, the "reversal" design, and the "multiple-baseline."

In the ABA design, the pupil's performances are first assessed for several days under a "no-treatment" condition (the first "A" phase). After a reasonable baseline or estimate of the child's pretreatment performances and rate of progress has been established, treatment is begun (the "B" phase). After several assessments in the treatment phase, the treatment is removed and several non-treatment assessments are collected once more (the second "A" phase). If the treatment is effective, then performances will shift markedly in the desired direction during the "B" phase and then drop again during the second "A" phase. If performances do not drop during the second "A" condition, then one must conclude that the increase in performance during the "B" condition might only be due to some event which coincidentally occurred at the time the special program was begun.

The ABA design can, of course, be extended to include many applications and withdrawals of the program and/or to test a series of different programs (e.g., ABACABAD...), with each new shift in performance (up or down) adding to the evidence of program effectiveness. Unfortunately, such a design must assume that the treatment or program is prosthetic--that higher or more desirable levels of performance will only continue to be observed for as long as the program is in effect. Most

educational programs are by design therapeutic, however, and are specifically intended to produce changes in behavior which will continue to be observed in the pupil after the program is terminated. If that is the case, then no reduction in performance would be expected during the second "A" condition and the design is inappropriate.

"Reversal" designs are devised to produce shifts in performance similar to those desired with the ABA design, but they are much more powerful in doing so. After a baseline of preprogram performances has been established, the program is introduced, just as in the ABA design. Following a series of assessments of the program's influence on performances, however, the program is not simply terminated. Rather, the program conditions are systematically altered to deliberately produce the opposite of the normally intended behavior change. For example, if the program was originally designed to reinforce high reading rates, the reinforcement contingencies are changed to deliberately encourage low reading rates. If reinforcement is, indeed, effective in determining reading rates, then reading rates should go down. Such designs are extremely powerful in demonstrating the effects of program elements, but they are also somewhat questionable from an ethical standpoint. There are only a few situations in which one could justify the active encouragement of undesirable behavior simply for the sake of demonstrating the power of the same procedures in developing desirable behavior. In most cases, the reversal design offers little for the applied educational practitioner.

Multiple-baseline studies manage to avoid most of the problems associated with either ABA or reversal designs. Rather than simply measure one behavior or pupil, several behaviors or pupils are studied. After reliable baselines of pupil performances on the several programs have been established, one of the programs is changed. Later, a second program is changed, still later a third, and so on until all the programs have been shifted from the pre-treatment conditions to the treatment conditions. If the program is effective, then pupil performances will shift when and only when the program is introduced. If shifts in performance are due to some other temporally related event, then shifts will occur in unison, not just when the program is introduced to each child or with each successive behavior. Since temporal confounds can be controlled by staggering the time when the program is introduced, there is no reason to remove the program or to institute special procedures designed to produce a drop in performances. Many of the practical and ethical concerns encountered with ABA or reversal designs are thus circumvented. Still, multiple baseline approaches to evaluation do not solve all of the problems in program evaluation.

Establishing the Baseline

The validity of multiple baseline studies rests, in large part, on the manner in which baseline or pretreatment performances are assessed. Unless some confidence can be gained that preprogram performances have been meaningfully and reliably determined, the rest of the evaluation will be meaningless. Establishing such a baseline in many applied situations may prove somewhat difficult.

Withholding treatment. The most common approach to establishing preporgram baselines is simply to withhold treatment from the pupil for some period of time. To some extent, that presents the same ethical concerns encountered when trying to justify a "no-treatment" control group. If baselines can be established in a matter of a few days or weeks, however, that problem may not be of real concern. There are even two reasons why a program should not or cannot be started immediately: teachers cannot be expected to begin all programs with all children on all behaviors at once; and in at least some cases, a child might not actually be ready for a particular program. If so, then the "natural" delay in program implementation might be used to establish reliable baselines. It's not quite that simple.

If a teacher cannot begin all programs at once, then there is some doubt as to whether he or she would have the time to collect the baseline information for those programs. Even if information can be collected, there still remains the problem of comparing baseline performance data which were collected under conditions when absolutely no program was in effect with performance data collected under program conditions. Most educators would not be impressed by the fact that when you didn't work with a behavior at all, the child performed very poorly, and then when something (i.e., anything) was done to develop the behavior, the child improved. So what? Comparing something against nothing hardly demonstrates the program's superiority over alternative program options.

Similarly, if a program were withheld because a child was not yet ready for it, then baseline measures collected during that period could hardly be considered a meaningful basis of comparison for performance measures collected during the program. Perhaps "readiness" alone accounts for any observed increases in performance.

It seems apparent, therefore, that "natural excuses" for withholding treatment during baseline periods can create more problems than they solve. Either the educator must face the ethical issues of withholding treatment for some (hopefully brief period of time), or some other means of estimating preporgram performances must be found.

Historical Data

In some cases, preprogram baseline performances can be established by analyzing performance data collected prior to program referral and/or by utilizing data collected during the referral process itself. The availability of meaningful pre-referral data is generally rather low however, and even when data are available, their reliability is frequently questionable.

In light of those problems, many evaluators have attempted to estimate preporgram rates of progress based on the child's age and level of performance at the time he or she enters the program. Assume, for example, that a program is about to be implemented with a child who should have already fully developed the skill in question by the time of her referral to the program. At the time of entry, child is performing the skill at only 30% of "mastery," and one year has passed since the

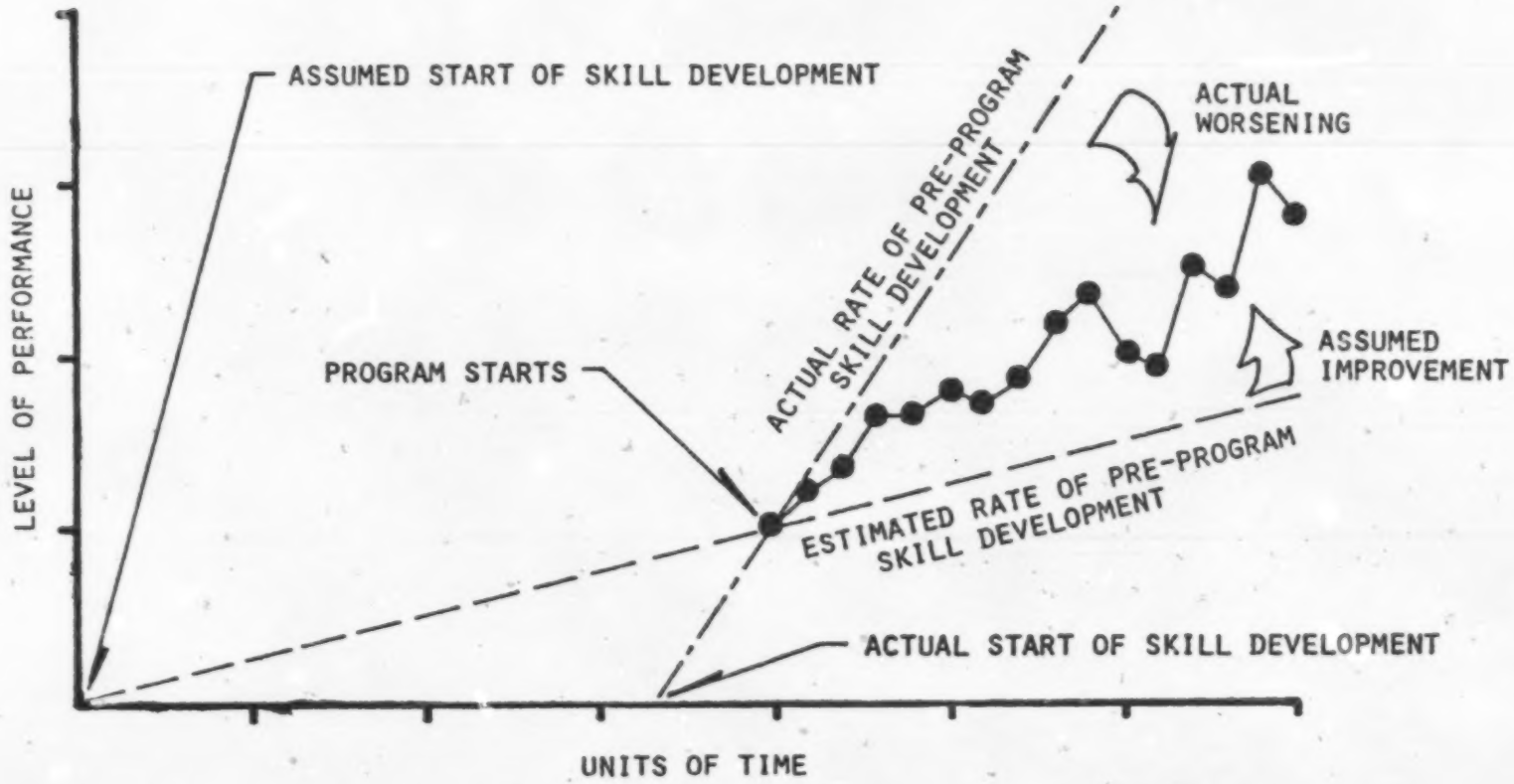

Figure 3: The effects of erroneous assumptions concerning the point when skill development began in estimating pre-program rates of progress.

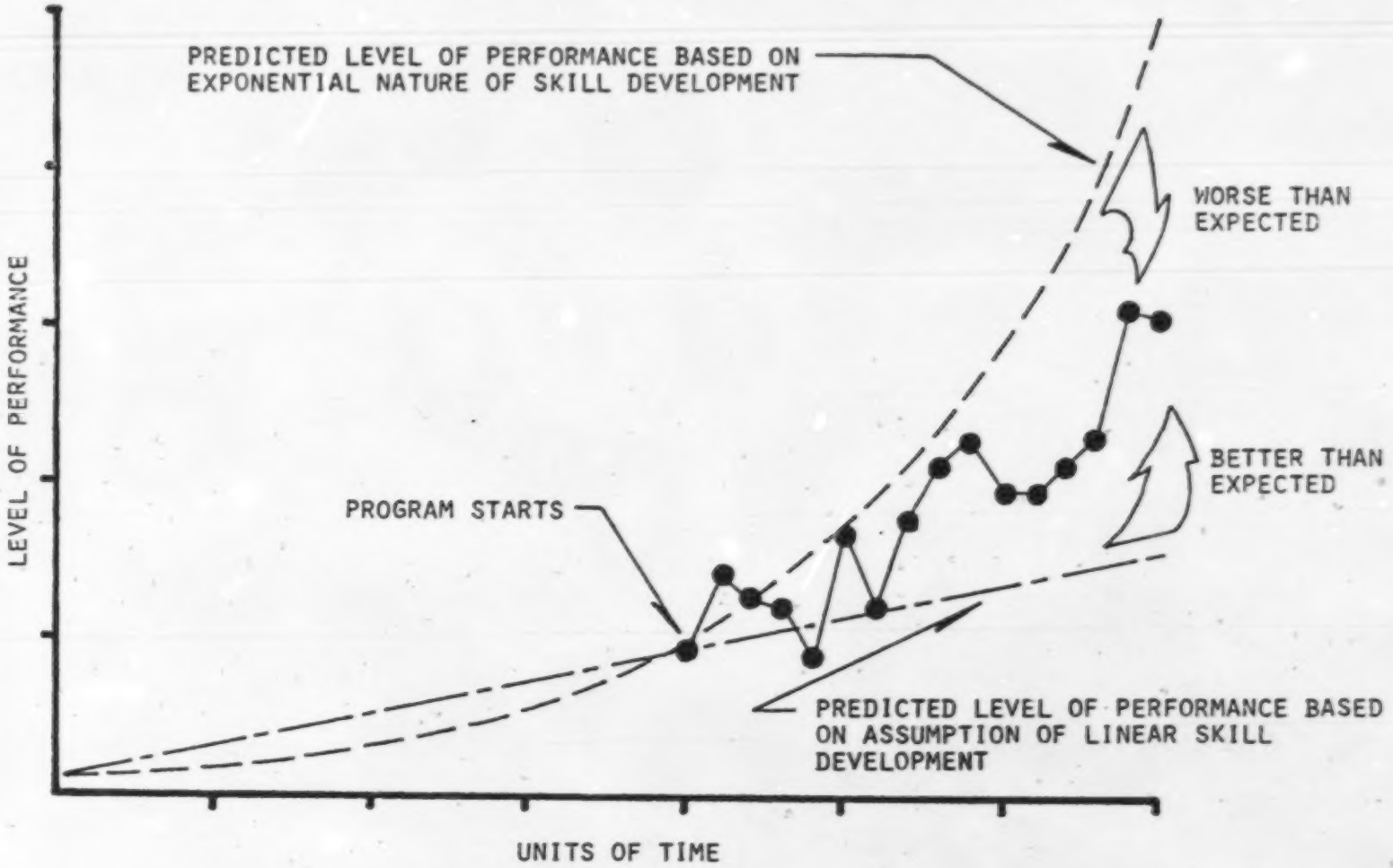
PREDICTED LEVEL OF PERFORMANCE BASED ON EXPONENTIAL NATURE OF SKILL DEVELOPMENT
WORSE THAN EXPECTED
BETTER THAN EXPECTED
PROGRAM STARTS
PREDICTED LEVEL OF PERFORMANCE BASED ON ASSUMPTION OF LINEAR SKILL DEVELOPMENT
LEVEL OF PERFORMANCE
UNITS OF TIME

Figure 4: Differences in evaluating program impact as a function of the type of skill development curve assumed.

skill would normally begin to develop in most children. The preprogram rate of development is assumed, therefore, to be (30% ÷ 12 months =) 2.5% per month. During the course of the program the child develops at a rate of 12% per month, so the program is believed to have accelerated her development. Obviously, a great many assumptions have been made.

First, one must assume that the child actually began to develop the skill at the "normal" or expected time. That assumption appears particularly tenuous in light of the fact that the child was delayed in development at the time of program implementation. The child might actually have started to develop the skill one month prior to referral, in which case the preprogram rate of development was more than twice that which was observed during the program (see Figure 3).

- - - - - - - - - - - - - - - -
Insert Figure 3 about here
- - - - - - - - - - - - - - - -

Secondly, many skills are acquired in an exponential fashion--each new step in skill mastery is easier to make than preceding steps. If that is the case, then the absolute rate of growth for any period later in the development of the skill would be greater than any preceding period, with or without the program (see Figure 4). The application of time-series approaches to the evaluation of skill development (i.e., the use of repeated assessments at several points in time) should minimize this type of analytic error, since the growth record during the program should point out any non-linearities in skill development. Nevertheless, the potential for misinterpretation of such records has been realized on numerous occasions.

- - - - - - - - - - - - - - - -
Insert Figure 4 about here
- - - - - - - - - - - - - - - -

Several methods for obtaining reasonable estimates of preprogram progress have been suggested (e.g., Brck & Weisburg, 1977; White & Haring, in press), but all depend in some part on group analyses of the relationships between the rate of skill development, entry performance level, and age. For such analyses, of course, a common metric must be found which enables the results of "individualized" programs to be collated across pupils--a problem which I will address somewhat later in this presentation.

Treatment vs. Treatment

As already pointed out, most applied educational programs are in a constant state of some change. As indiviudal or group programs fail to produce the expected results, programs are modified. If time series data are being collected, it is sometimes possible to document the impact of those changes with some precision. Events which appear to have a facilitating impact on one program might then be tried with other pupils or programs, thereby creating a natural multiple baseline design which avoids the "withholding of treatment" problem. Of course, such a

procedure amounts to little more than capitalizing on one's failures--demonstrating that what one thought would work did not, and that some alternative program did. As the overall program improves, failures will decrease, and the opportunity to employ such a strategy for overall program evaluation is reduced. It's a nice approach, but with limited application.

What About the Overall Picture?

While not without problems, multiple baseline designs for the evaluation of individual educational programs do offer some promise. When properly conducted, they can provide very convincing evidence of a program's impact and (in the case of multiple baseline studies across subjects), the generalizability of a program as well. Perhaps just as importantly, such data are far more useful to educators in making timely and appropriate decisions about a child's program than the infrequent type of testing schedule typically associated with group evaluation designs. Due to the several constraints and problems mentioned earlier, however, it is not likely that all of a child's program will be amenable to such investigations, and perhaps there will be some children in a program for whom even a single well constructed multiple baseline study is difficult. That raises the issue of utilizing such information in evaluating a program's overall effectiveness in meeting the needs of all its pupils in all areas of concern. Namely, are those situations in which multiple baseline studies can be conducted representative of the program as a whole? If not, then how can those studies be placed in some perspective? To answer those questions, we must begin to search once more for a common metric which can be meaningfully applied to all children within the program.

Curriculum-Referenced Assessments

The absence of meaningful nationally validated norms, many programs serving special subgroups of pupils have turned to what has been called "curriculum-referenced assessments." According to English and English (1958), an assessment instrument has curricular validity to the degree "test items are representative of the knowledge, skills, or attitudes supposed to be learned in a given course of study" (p. 575). In order to construct a valid measure of a child's progress through a program, therefore, perhaps a simple list could be compiled of the program's overall objectives, and then checked-off as they are mastered. To the extent that all children in a given program share at least some needs in common, they might all be meaningfully assessed with such an instrument. Moreover, what better measure of a program's impact than a metric which reflects whether children have learned what the program has attempted to teach? It is not quite that simple, of course. Several factors must be carefully considered.

Scaling

If a child's score on a curriculum-referenced checklist is to have any meaning, several basic scale requirements beyond smiple measurement

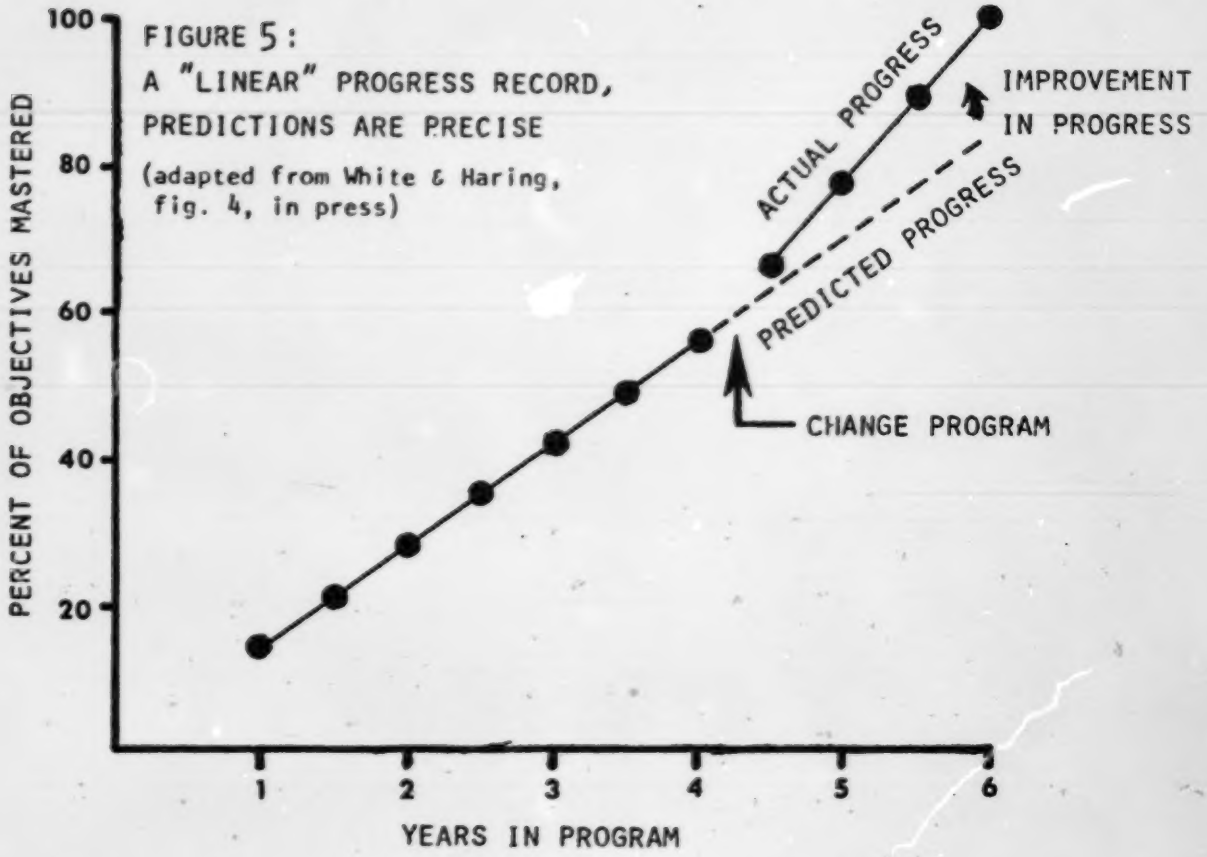

FIGURE 5:
A "LINEAR" PROGRESS RECORD,
PREDICTIONS ARE PRECISE
(adapted from White & Haring, fig. 4, in press)

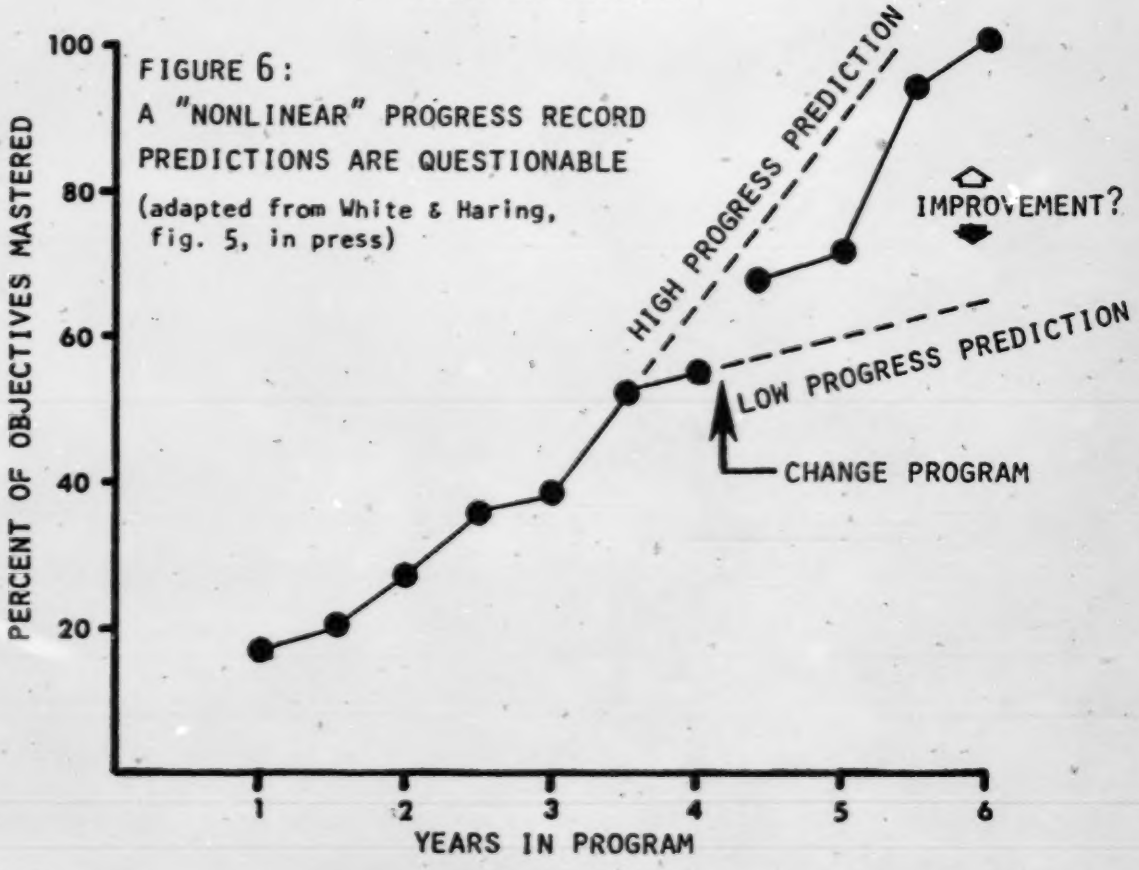

FIGURE 6:
A "NONLINEAR" PROGRESS RECORD
PREDICTIONS ARE QUESTIONABLE
(adapted from White & Haring, fig. 5, in press)

reliability and content validity must be met. First, one must have confidence that the scale is constructed in a manner that allows at least ordinal decisions to be reached--it should be possible to assume that given two children with different scores, the child with the higher score has, indeed, reached a higher level of competence. If objectives are always mastered in a particular sequence (regardless of the method of instruction), then ordinal measures are assured. If, however, it is possible for children to master objectives in more than one sequence, then objective weighting becomes an issue. Each objective must represent the same amount of "functional" learning or developmental increment as every other objective. Only in that way can a child with a higher score than another be assumed to have actually developed to a higher (albeit, perhaps qualitatively different) level. If the developmental significance of different objectives can be assumed (either by the way in which they are written or through some system of differential weighting), then an additional advantage will have been gained. It will not only be possible to compare two or more children (as, for example, might be the case when a contrast group is available), but it will also be possible to use the curriculum metric itself in time-series evaluations of individual or group progress (see Figures 5 and 6).

- - - - - - - - - - - - - - - - - - -
Insert Figures 5 and 6 about here
- - - - - - - - - - - - - - - - - - -

Do All the Objectives Apply?

Even if appropriate scaling of objectives is achieved, one must question whether the metric can be meaningfully applied to all children within the program. After all, many of the arguments presetned earlier in this presentation have been directed at the issue of "individualized" programs. Is it reasonable to assume that a single list of objectives could serve equally well the needs of all individuals? Perhaps not. One must bear in mind, however, that the objectives in question should at least be "program-relevant," and that all children in the program should share at least some needs and objectives in common. That is not at all equivalent to saying that all the skills measured by a certain nationally normed intelligence or ahcievement test are applicable to the children in a particular program. Care must still be taken to specifiy objectives in the most individually adaptable fashion possible (White, Reavis, & Kunzelmann, 1978) and some compromise will be necessary. Not all objectives will be equally important for all pupils, but all objectives should still have some relevance to the desired or expected progress of such children. Indeed, even though not all children will be provided instruction on each objective, it is still desirable to assess their progress with respect to as broad a spectrum of skills as possible. As Tallmadge (1977) pointed out:

> It is occasionally possible to conduct a credible evaluation without an empirically based estimate of what conditions would have been without the intervention. Where a totally new topic is introduced into a curriculum, for example, it is not necessary to demonstrate empirically that a control group, not exposed to the topic,

learned less of it. If the treatment group learned anything as a reuslt of the new instruction, it could be clearly expected to outperform the control group and no proof would be required. On the other hand, it should be demonstrated that the new learning was in addition to, not in place of, what was learned previously (p. 18).

It is important, therefore, that any curriculum-referenced instrument not be so restricted to the particular needs of each child that overall development is ignored. Although instruction may not be provided in all areas of a child's development in a given program, all important areas of development should still be sampled. If learning is demonstrated in those areas toward which program efforts are most heavily directed, but learning falls off in other important areas, then the actual success of the program must be questioned.

It Can Be Done!

Despite the obvious difficulties in devising a reasonable curriculum-referenced instrument, it is possible. Several such instruments already exist for programs serving moderately, severely, and profoundly handicapped children (e.g., the University of Washington's Uniform Performance Assessment System; the State of Oregon's Student Progress Record; the Teaching Research Curriculum)[1], and a new instrument is currently being developed by an interstate consortium of programs (the Consortium on Adaptive Performance Evaluation)[2] which promises to be particularly adaptive to the unique needs of multiply handicapped children. Not all available instruments are equally well suited to a given program, of course, but many instruments are accompanied by computer backup software which enables the precise analysis of the impact of any additions, deletions, or revisions of the scale which one might wish to make. Indeed, some computer programs allow users to test the scale characteristics of an entirely new curriculum-referenced instrument devised to meet the special needs of any particular program (e.g., the Uniform Performance Assessment System, Bendersky, 1977).

---

[1]A comprehensive listing of such instruments has been compiled by Lestra Hazel and is contained in the Experimental Education Training Program, Norris Haring, Editor, available from the Experimental Education Unit, CDMRC, WJ-10, University of Washington, Seattle, WA 98195.

[2]For additional information concerning the Consortium on Adaptive Performance Evaluation, contact Dr. Dale Gentry, College of Education, Department of Special Education, University of Idaho, Moscow, Idaho 83843.

The Value of Curriculum-Referenced Instruments

Using a curriculum-referenced instrument alleviates many of the problems associated with subject-unique time series data. First, since a well constructed curriculum referenced instrument will assess progress in both programmed and nonprogrammed areas, it is possible to evaluate a child's progress from a borad perspective-thereby reducing the probability of a "so what?" reaction to demonstrated gains. Secondly, since the performances of all children in the program are held relative to the same set of objectives, group summaries and analyses are possible. Such summaries can be most useful when making overall program development decisions. Third, since a long-term sequence of skill development is assessed, and since it is possible to analyze progress records across children, certain types of extrapolation procedures can be more easily justified. The problems with extrapolation mentioned earlier centered around the estimation of the time when skill development might reasonably be expected to have begun, possible counfounds with age and level of deficit, and problems in determining whether skill development was linear or curvilinear. By measuring a sequence of skills, it is possible to develop a sequence which reaches back to some fairly well defined opportunity for the initation of skill development. With a preschool scale, for example, the developmental sequence can be extended back to the birth of the child. Linearity of expected development can be assumed if the scale was properly constructed and tested, and even retested longitudinally after the child enters the program and/or by a cross-sectional analysis of those children already in the program. Possible confounds with age and level of deficit can also be tested cross-sectionally by analyzing the relationship between rate of progress through the curriculum and age and/or level of deficit for children already in the program. In many cases, therefore, it should be possible to estimate meaningful and believable baselines against which further development can be compared (e.g., Haring, 1978; White & Haring, in press).

Additional Benefits

As menthioed earlier, no ongoing system of evaluation is likely to survive if it does not provide some immediate benefit to the educators who must make it work. Fortunately, the curriculum-referenced approach offers benefits to teachers and administrators alike:

(1) Program goals are more clearly defined by a curriculum-referenced instrument than by almost any other description of program intent conceivable. It is easier, therefore, for funding agencies to determine the program's potential merit, for referring agencies and parents to determine whether the program might represent a reasonable placement option for a particular child, and for teachers to decide exactly where their instructional emphasis should be placed.

(2) Greater program balance can be achieved if the curriculum-referenced instrument is itself well-balanced in all major areas of developmental concern. Teachers need not actually instruct each child in all areas of development, they can concentrate on what they feel are the child's greatest needs.

If such an emphasis is appropriate, then the child will continue to develop in all areas (as measured by the curriculum referenced instrument). If, on the other hand, overemphasis of a subject or skill area is detrimental to the child's continued progress in other areas, that will be easily determined.

(3) Continued refinement of the curriculum is facilitated by noting areas in which most children in the program have special difficulties. Sequence and linearity analyses will also help to pinpoint areas in the curriculum which might best be rearranged, restructured, or refined in some way.

(4) The development of Individualized Educational Programs (IEP's) is facilitated by providing a ready-made list of the most common objectives which might be chosen for a child. Records of which objectives are most typically selected, which are least often selected, and which are usually modified extensively in some way, will also provide additional information for changes in the basic curriculum and the curriculum-referenced instrument itself.

(5) The compilation of longitudinal time-series data which is essentially comparable across children and subprograms provides the basis for the analysis of major changes in the basic program. Of course, the precision of such analyses will depend, in part, on the frequency with which data are collected and the frequency of changes in the program to be analyzed. With many programs, where many of the objectives are based on the expectancies of children in alternative program placements, there is even a relatively high probability that actual preprogram and/or postprogram data might be available for more precise time-series analyses of program impact.

Still No Panacea

Curriculum-referenced approaches to evaluation alleviate some problems, but they are still a long way from being a panacea. They avoid some of the difficulties in applying national norms to special subgroups in the population, but only at the expense of providing more ideosyncractic data of less use in cross-program comparisons. Similarly, they enable the overall progress of children in a given program to be assessed with a common metric, but only by ignoring at least some of the important individual differences between children. Devising objective lists which can be applied to more programs aggravates the problem of accounting for individual differences, and greater specialization to account for individual differences makes the metric less useful for cross-program evaluation. Striking a reasonable balance can be difficult, and the process of validating an instrument once the basic issues of its design have been resolved can be a long and arduous task. Nevertheless, the curriculum-referenced approach does appear to offer at

least a reasonable chance of pulling together individual time-series data for overall program evaluation in a way that is useful to educators. Serious consideration should be given to its use and further development as an alternative or adjunct to more traditional group or single-subject program evaluation strategies.

References

Bendersky, M. Uniform Performance Assessment Working paper No. 65, Revised March, 1977.

Bryk, A. S. & Weisberg, H. I. (Co-principal investigators). Evaluation Methodology for Early Childhood Special Education Programs. Grant application submitted to the U.S. Office of Education (Handicapped), 10 Jan. 1977, Huron Institute, Cambridge, Massachusetts.

Campbell, D. T., Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963. (Also published as Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.)

Cooley, W. W., & Lohnes, P. R. Evaluation research in education: Theory, principles and practices. New York: John Wiley and Sons, 1976.

English, H. B., & English, A. C. A Comprehensive Dictionary of Psychological and Psychoanalytical Terms. David McKay Co., Inc., New York, 1958.

Haring, N. G. Center for Severely Handicapped, report to the Joint Dissemination and Review Panel, 1978. Unpublished manuscript. Available through: Experimental Education Unit, University of Washington, Seattle, WA 98195.

Hersen, M., & Barlow, D. H. Single Case Experimental Designs: Strategies for Studying Behavior Change. New York, Pergamon Press, 1976.

Horst, D. P., & Tallmadge, K., & Wood, C. A Practical Guide to Measuring Project Impact on Student Achievement. RMC Research Corporation, Mountain View, California, 1975.

Sailor, W., & Mix, B. J. TARC assessment system. Lawrence, Kansas: H & H Enterprises, 1975.

Sidman, M., Tactics of scientific research: Evaluating experimental data in psychology. New York: Basic Books, 1960.

Tallmadge, G. K., & Horst, D. P., A procedural guide for validating achievement gains in educational projects. Mountain View, California: RMC Research Corporation, 1976.

Tallmadge, G. K. The Joint Dissemination Review Panel, Ideabook, RMC Research Corporation, Mountain View, California, 1977.

White, O. R., Reavis, K., & Kunzelmann, H. Child Performance Based Decisions for Screening, Placement, and Adaptive Program Planning. Presentation made for CEC Topical Conference on Individualized Educational Plans, Albuquerque, February, 1978.

White, O. R., & Haring, N. G. Evaluating Educational Programs Serving the Severely and Profoundly Handicapped, Teaching the Severely Handicapped: A yearly publication of the American Association for the Education of Severely/Profoundly Handicapped, Vol. 3, Columbus, Ohio, 1978 in press.